Research Paper | Information Technology | E-ISSN No : 2454-9916 | Volume : 2 | Issue : 4 | April 2016

SUMMARIZING DATA BY USING DATA MINING TECHNIQUES A COMPARATIVE BY USING C4.5 AND C5.0 ALGORITHMS

Nima Abdullah AL-Fakhry

College of Administration & Economic, Department of MIS ,University of Mosul

ABSTRACT

This paper describes the theoretical assues of the data mining concept and their development steps. The differences between clustering and classification process are identified. The practical sides of C4.5 and C5.0 classifiers one dealt with and a comparative study is held. The results of applying both classifiers on 30 patients test serums are illustrated and compared via both classifiers. The comparison highlights the superiority of the C4.5 algorithm in presenting high resolution.

KEYWORDS: Data mining, Machine Learning, C4.5, C5.0,Gain ratio.

1. Introduction:

During the widespread use of information technologies of the most significant features that have been applies in many aspects of life, new challenges have emerged in the field of databases utilize such features are not as warehouses for data only, but also for searching the information extracting of knowledge bases. Since the mid-sixties of the last century, the works began in employing of algorithms, in exploration, in evaluating strength properties of each algorithm. The works also evolve and deriving models, combining the properties, using the method base correlation and clustering, and then using all these models in various fields ranging from genetic algorithms, probabilities conditions, construction of future predictions, and exploring the behavior, allowing trends to takethe right decisions and taking in a timely manner.

Therefore the existence of information systems has become an urgent necessity in order to deal with the data and information in terms of storage, retrieval, display and use them in decision-making and planning.

That why a new branch of science of artificial intelligence appeared. It is a science of data mining, which aims to build same models. These models are algorithms that connects a set of inputs to obtain the knowledge of the task and the new data after arranged and organized upon common characteristics among them while taking the advantages of them for interpreting the obtained results.

Accordingly, it stems from the hypothesis that, the C4.5 algorithm is better than theC5.0 algorithm in data mining based on laboratory tests conducted by heart patients.

In this paper, a theoretical frame work for the concept of data mining is presented. In addition, C4.5 and C5.0 are described and then compared based on performance evaluation and percentage of gain information.

This paper is organized as follows: In addition to this introductory section, section 2 gives some theoretical background. Experimental results are contained in section 3. In section 4, a comparative study is pointed based on some statistical analysis. Finally, Section 5 concludes this paper.

2. The theoretical aspect:

A. The concept of data mining:

Data mining is the process of using some techniques of statistical ,mathematical ,…, etc to identify and extract some useful information and new knowledge from databases or data warehouses. This means a search for hidden regular knowledge in a large number of incomplete data which isconfusing ,mysterious , random , and it is un known in advance to users. In spite of that , at the end it be understood as information, useful and practical knowledge. It means Advanced knowledge is unexpected prior information , or updated information , in which information discovered is more surprising , more likely to be actually effective in future. This information or knowledge are being effective, practical , and achievable by some algorithms.

Data mining is associated closely with discovering knowledge which is a multidisciplinary science, has a database , integrated information, and it is one of the modern techniques of artificial intelligence, machine learning and statistics. Databases, artificial intelligence, and data mining, are the most important categories of the three pillars of the big powerful technology in the current era(Zheng,2012).

The objective of data mining is to extract hidden predictive information from large data bases. It is a powerful technology with great potential to assist organizations and institutions to focus and direct its sights on the most important information in their data warehouses. Data mining tools help predicating future trends and behaviors. They help institutions and organizations to make decisions driven by prior knowledge of the mechanism, and provide tools exploration analyzes of past events retroactively. These tools can also answer questions that take longer than necessary to solve them, and help in finding predictive information that has been absent from the minds of experts because it lies outside their expectations (Gupta & Todwal,2012).

B. Goals of data mining:

Mining the data bases aims to extract hidden information, It is a modern technology that has become important under the rapid development and wide spread use of data bases and competition in the markets and others. Their use provides for institutions in all areas, the ability to explore and focus on the most important information in data bases. Mining techniques is building the future predictions and can explore the behavior and trends, allowing an estimate to take proper decisions at the appropriate time.

Mining techniques can answer many questions in standard time, especially those who are difficult to answer, if not impossible, using traditional statistical techniques , and those questions which take a long time and many of the analysis procedures to solve (Nisbet and et al, 2009).

C. Reasons for the evolution of applications of data mining:

Data mining applications has started grow significantly for the following reasons(Adrian & Zantinge 2010):

1. The amount of data in the data store and data market is growing very significantly, because of the presence of a large IT environment push those interested to take advantage of them.

2. The emergence of many effective miningtools,has encouraged the increase in mining operations in the data frequently.

3. Intense competition in the markets paid companies to look for ways to assist them successfully in minimal costs.

D. Differences between clustering and classification:

The following table illustrate the differences between clustering and classification (www.broadinstitute.org) (Shuweihdi,2009):

Table(1)
Difference between clustering & classification

	Clustering		Classification
1	it's points are not described	1	Some of it's points are described
2	The clusters are based on the close between data sets.	2	It needs a law or rule based on it's accurately.
3	It's one of un supervised machine learning techniques	3	It's one of supervised machine learning techniques
4	no predefined classification is required. The task is to learn a classification from the data	4	the task is to learn to assign instances to predefined classes (keller,2001)

E. Data mining in Machine learning :

Machine learning is an area of artificial intelligence concerned with the study of computer algorithms that can be improved automatically through experience. In practice, this involves creating programs that optimize a performance criterion through the analysis of data. The Types of machine learning are (Sewell,2007):

1. Supervised learning : The algorithm is first presented with training data which consists of examples which include both the inputs and the desired outputs, thus enabling it to learn a function. The learner should then be able to generalize from the presented data to unseen examples, C4.5, C5.0, and CART algorithms are examples working in environment of supervised learning.

2. Unsupervised learning: The algorithm is presented with examples from the input space only and a model is fit to these observations. For example, a clustering algorithm would be a form of unsupervised learning K-means and Auto class algorithms are examples of working in environment of unsupervised learning.

3. Reinforcement learning: An agent explores an environment and at the end receives a reward, which may be either positive or negative. In effect, the agent is told whether he was right or wrong, but is not told how.

F. C4.5 & C5.0 Algorithms:

1. C4.5 Algorithm:

This algorithm is an improvement of ID3. It can work with numerical input attributes as well. It follows three steps during tree growth:

1. Splits creation for categorical attributes is the same as in ID3. For numerical attributes all possible binary splits have to be considered. Numerical attributes splits are always binary.

2. Evaluation of best split for tree branching based on gain ratio measure, and

3. Checking of the stop criteria, and recursively applying the steps to new branches.

This algorithm introduces a new, less biased, split evaluation measure (Gain ratio). The algorithm can work with missing values. It has pruning option, grouping attribute values, rules generating etc (Suknovic & et al, 2012)(Dai and Ji,2014).

The Gain ratio selection criterion is a measure that is less biased towards selecting attributes with more categories (Hamilton & et al, 2012) (Adhatrao and et al, 2013)

Where P, n representing different varieties.

In the case of a particular attribute for example (A) , and k has a different values, so the decision tree formula is:

$$E(A,p,n)=\sum_{i=1}^{k}\frac{p_i+n_i}{p+n}I(p_i,n_i)...(2)$$

Where n_i, pi represent the numbers of cases for each class from the decision tree and connected with the part I depending on the value of A. The final formula for gain is given by:

$$Gain\ (A,p,n)=I(p,n)-E(A,p,n)...(3)$$

2. C5.0 Algorithm:

C5.0 model works by splitting the sample based on the field that provides the maximum information gain. The C5.0 model can split samples on basis of the biggest information gain field. The sample subset that is driven from the former split will be split afterward. The process will continue until the sample subset cannot be split any more. Finally, bt examining the lowest level split, those sample subsets that don't have remarkable contribution to the model, will be rejected (Patil & et al, 2012)

C5.0 constructs models for classification by using inductive, supervised machine learning. Input consists of a set of training items, each of which is described by a single record consisting of attribute-value pairs. Each item in the training set is assigning one of a predefined set of discrete classes (this is supervised learning) (Bankert & et al, 2004) (Pandya, & Pandya, 2015).

The formula of gain ratio is computed from the following equations:

$$\text{Gain(s,v)} = E(s)-\sum_{v}\frac{|Si|}{|S|}*E(Sv)...(4)$$

$$\text{SplitInfo(s,v)} = \uparrow_{i=1}^{m}-\frac{|Si|}{|S|}*\log 2\frac{|Si|}{|S|}...(5)$$

$$\text{GainRatio(s,v)} = \frac{Gain(s,v)}{SplitInfo(s,v)}...(6)$$

3. Experimental results:

Data mining technique was designed to extract knowledge from large amounts of data processing and take appropriate decisions. In this section, the practical side of the adoption category by C4.5 and C5.0 algorithms is presented and then compared to determine the most appropriate algorithm to obtain information on tests of laboratory conducted by heart patients.

The experiment data include tests serums: sodium, potassium, magnesium, chloride, calcium, phosphorus, creatinine, Total S. Cholesterol, triglycerides, the percentage of urea in the blood, and uric acid for 30 Patients, (15) Male and (15) Female (Appendix (1)),Fig.(1) shows the flow diagram for the tasks:

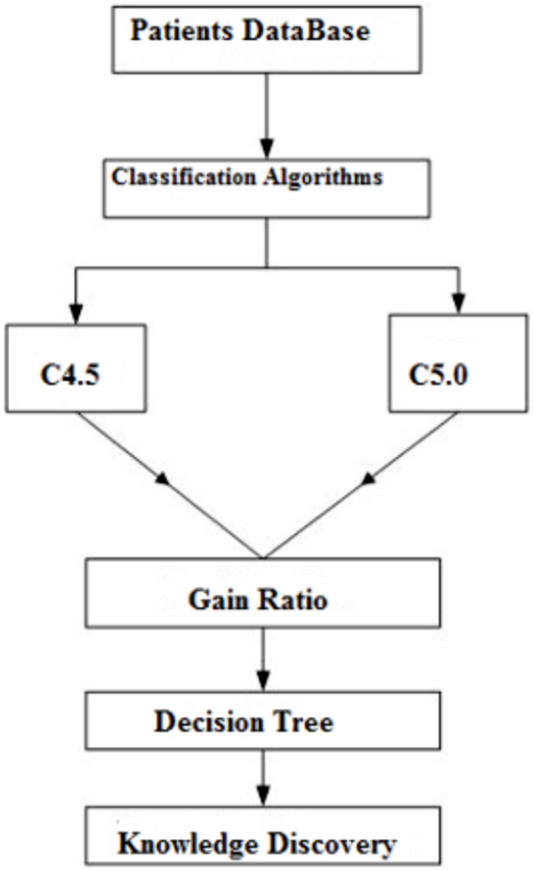

Figure (1)
Experiments flow Diagram

At the first stage, C4.5 is used as classifier and the equations (1),(2)&(3)are applied to obtain the Gain Ration for each test, Table (2)shows the results and Figure (2) shows the ratios of serums.It can be seen that the greatest is one (Total S. Cholesterol),the red part :

Table(2)
The Gain Ratio for tests / C4.5

Sequence	Gain Ration	Serums
1.	856.042764	Total S. Cholesterol
2.	720.161263	Calcium
3.	673.87304	Triglycerides
4.	574.4058969	Potassium
5.	521.0822333	Serum creatinine
6.	240.10556	Serum chloride
7.	165.7288615	Serum sodium
8.	157.5520106	Blood Urea
9.	95.302695	Phosphorus
10.	39.998249	Magnesium
11.	4.901793	Uric acid

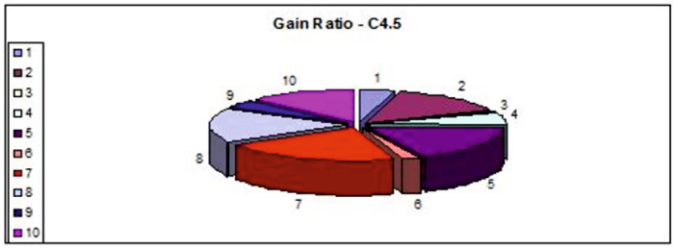

Figure (2)
Gain Ratios of C4.5 Algorithm

From Table (2) and Figure (2), the highest gain ratio obtained is blood lipids (cholesterol) and the lowest gain ration from (uric acid).Thisstep is important to know the test with as mall percentage, because it is not necessary to be adopted when evaluating the status of the patient, the results of the algorithm are shown in Figure (3) and Table (3).

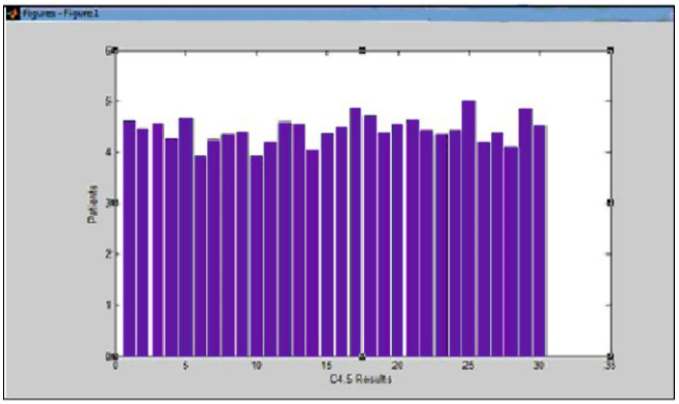

Figure (3)
The Result of C4.5 algorithm

At the second stage, use the C5.0 classifier and apply the equations (4),(5),(6) to obtain the Gain Ration for each test, table (4) show the results and figure (4) show the ratios of serums and the greatest one (Total S. Cholesterol) the red part :

Table (3)
The Results of C4.5 algorithm

Sequence	Sex	Results
1.	M	138.0397487
2.	M	133.5156172
3.	M	137.0334029
4.	M	127.5960163
5.	M	139.8487279
6.	M	117.4128112
7.	M	126.9566652
8.	M	130.4756757
9.	M	131.6837125
10.	M	117.4807367
11.	M	126.3617824
12.	M	137.7178279
13.	M	136.8514331
14.	M	121.5230214
15.	M	130.8474924
16.	F	134.0923411
17.	F	145.8919
18.	F	141.6147896
19.	F	131.22635
20.	F	136.8648337
21.	F	138.5788739
22.	F	132.7465243
23.	F	130.5612072
24.	F	132.8867456
25.	F	150.0631171
26.	F	125.9728886
27.	F	131.3212418
28.	F	122.6869135
29.	F	143.5628367
30.	F	135.5899698

Table(4)
The Gain Ratio for tests / C5.0

Sequence	Gain Ration	Serums
1.	7.55356	Total S. Cholesterol
2.	6.48508	Serum chloride
3.	5.52873	Serum sodium
4.	5.23646	Triglycerides
5.	1.52247	Blood Urea
6.	0.36035	Serum calcium
7.	0.26236	Uric acid
8.	0.17698	Serum potassium
9.	0.12324	Serum phosphorus
10.	0.04327	Serum creatinine
11.	0.03514	Serum magnesium

From the previous table, the highest gain ratio obtained is blood lipids (Cholesterol) as in C4.5 classifier and the lowest gain ration from (Serum magnesium),the results of the algorithm are shown in Table (5).

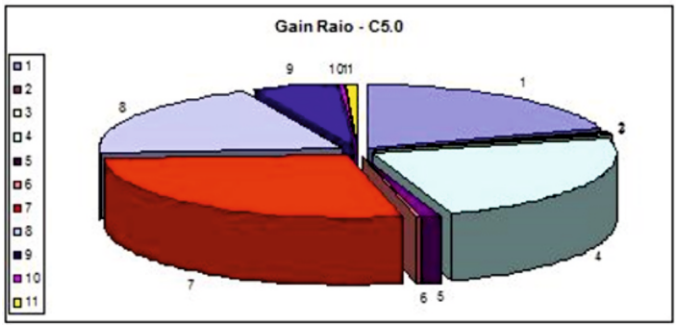

Figure (4)
Gain Ratio from C5.0 Algorithm

Table (5)
The Results of C5.0 algorithm

Sequence	Sex	Results
1.	M	0.9136651
2.	M	0.9136223
3.	M	0.9145298
4.	M	0.9107642
5.	M	0.9157549
6.	M	0.9054709
7.	M	0.9083931
8.	M	0.9117009
9.	M	0.9124724
10.	M	0.8996866
11	M	0.9087965
12	M	0.9094627
13	M	0.9087187
14	M	0.9078844
15	M	0.9110046
16	F	0.9115576
17	F	0.9157891
18	F	0.9136055
19	F	0.9095756
20	F	0.9124258
21.	F	0.9132224
22	F	0.9132913
23.	F	0.9081293
24	F	0.9085828
25	F	0.9153903
26	F	0.9134308
27	F	0.9134644
28	F	0.9054999
29	F	0.9136787
30	F	0.9080584

From the previous table that the results of the implementation of the algorithm was close and the differences among them little, this is illustrated in figure (3), Whereas the results of the implementation of the algorithm C4.5 was more obvious, the results differentiated, and this clear in figure (2).

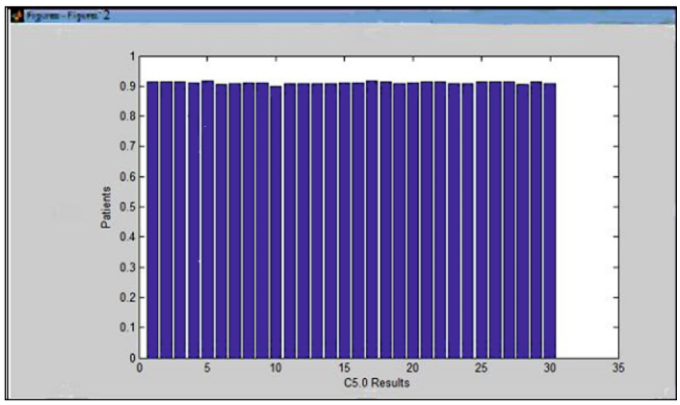

Figure (5)
The Result of C5.0 algorithm

4. A comparative Between C5.0 & C4.5:
Some statistical analysis one used to determine the performance of each of the used algorithms, These statistics include:

- **AVEDEV:** Returns the average of the absolute deviations of data points from their mean.
- **VAR:** Estimates variance based on a sample.
- **VARP:** Calculates variance based on the entire population.
- **STDEVP:** Calculates standard deviation based on the entire population.

Table (6) show this comparison:

Table (6)
Statistical Analysis for both classifier

Algorithm	AVEDEV	VAR	VARP	STDEVP
C4.5	5.910	59.604	57.617	7.590
C5.0	0.0028	1.29232E-05	1.24924E-05	0.00353

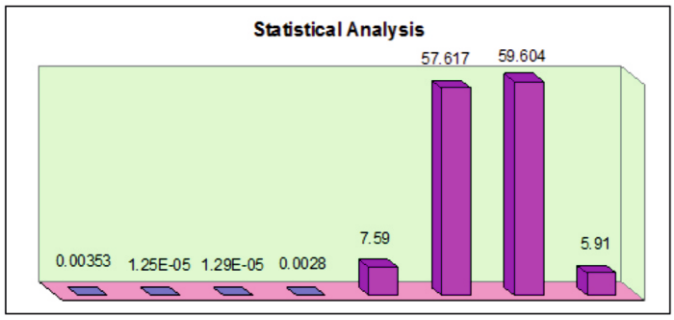

Figure (6)
Statistical Analysis for C4.5 & C5.0

4. Conclusion:
In this paper, a comparison between C4.5 and C5.0 classifiers data mining techniques has been presented using the heart patients data set. It has been shown that C5.0 algorithm needs less location at the application and less time execution. Alsoalgorithm C4.5has taken more storage space so it for needs more time for the implementation.Both algorithms can be represented either in a decision tree or in Aggregates of the rule sets, The gain ratio for C4.5 is higher than the gain in C5.0. From the statistical analysis, it has been observed that the performance of C4.5 algorithm is the best from the point of view of resolution , as shown in figure (6).

5. REFERENCES:
1. Researches:

1. Adhatrao, Kalpesh and Gaykar, Aditya and Dhawan,Amiraj and Jha, Rohit and Honrao, Vipul (2013),'PREDICTING STUDENTS PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS', International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5, September 2013, PP, 2,4.
2. BANKERT, RICHARD L., &HADJIMICHAEL, MICHAEL & KUCIAUSKAS, ARUNAS P., & THOMPSON, WILLIAM T & RICHARDSON, KIM (2004), 'Remote Cloud Ceiling Assessment Using data - mining Methods' ,Journal of Applied Meterology, Volume 43, P. 1935, Naval Research Laboratory Monterey, California, USA.
3. Dai, Wei and Ji, Wei(2014),"A MapReduce Implementation of C4.5 Decision Tree Algorithm", International Journal of Database Theory and Application Vol.7, No.1 ,P.4. http://dx.doi.org/10.14257/ijdta.2014.7.1.0
4. Gupta, Sita&Todwal, Vinod (2012), "Web Data Mining & Applications", International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249-8958, Volume-1, Issue-3,P. 20, February 2012.
5. Hamilton H. and Gurak, E. and Findlater, L. and Olive, W.(2012),"Overview of Decision Trees", Tutorials on decision trees, RudjerBoskovic Institute, P.3 University of California, California- USA, http://.machine-learning.martinsewell.com/machine-learning.pdf.
6. Keller ,Frank (2001), "Clustering-Connectionist and Statistical Language Processing", Computerlinguistik, Universit¨at des Saarla,P.1 Web site: www.coli.uni-saarland.de/~crocker/Teaching/Connectionist/lecture13_4up.pdf
7. Patil, Nilima&Lathi, Rekha&Chitre, Vidya (2012), "Comparison of C5.0 & CART Classification algorithms using pruning technique", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 4, June – 2012, Mumbai-India.
8. Pandya, Rutvija&Pandya, Jayati (2015), " C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning", International Journal of Computer Applications (0975 – 8887), Volume 117 – No. 16, P.2.
9. Sewell, Martin (2007) "Machine Learning", Department of Computer Science, P.1 University College London, Web Site:www.broadinstitute.org/annotation/winter-course_2006
10. Shuweihdi, Farag (2009)," Clustering and Classification with Shape Examples" Ph.Dtheseis, University of Leeds- Department of Statistics, P.2, UK. Web site: http://www1.maths.leeds.ac.uk/statistics.
11. Suknovic, Milija&Delibasic, Boris &Jovanovic, Milos&Vukicevic, Milan &Becejski-Vujaklija, Dragana, &Obradovic, Zoran (2012), "Reusable components in decision tree induction algorithms", Comput Stat (2012) 27:127–148, Springer-Verlag 2011.
12. Zheng, Yujie (2012), " Clustering Methods in Data Mining with its Applications in High Education", International Conference on Education Technology and Computer (ICETC2012), IPCSIT vol.43 (2012) © (2012) IACSIT Press, Singapore.

2. Books:

13. Nisbet, & Elder, John & Miner, Gary(2009) "Handbook of Statistical Analysis & Data Mining Applications", Academic Press/Elsevier, ISBN 9780123747655, P.12.
14. Adrian P. &Zantinge D.,(2010), "Data Mining",P.1, www.slideshare.net.

Appendix (1)

sex	Uric acid	Serum creatinine	Blood Urea	Triglycerides	Total S. Cholestrol	Serum phosphorus	Serum calcium	Serum chloride	Serum Magnesium	Serum potassuim	Serum sodium	Patient
M	6.9	1.4	40	130	211	2.9	8.7	158	0.88	4.1	137	x1
M	7	1.2	38	133	209	2.8	8.9	161	0.86	4.3	141	x2
M	6.7	1.1	37	122	200	3.1	8.9	170	0.9	4.4	148	x3
M	7.2	1.3	39	145	220	3	9.2	161	0.89	4.5	151	x4
M	6.4	0.9	34	120	195	2.7	9	164	0.91	4	145	x5
M	7.3	1.2	39	168	234	3.2	9.4	179	0.83	4.7	147	x6
M	6.2	1.6	44	152	225	2.5	8.5	181	0.88	5	151	x7
M	6.7	0.85	33	142	205	2.7	8.7	177	0.86	3.9	146	x8
M	7	1.1	37	137	200	3.1	9	188	0.93	4.2	143	x9
M	7.1	1.8	49	177	260	3.3	9.2	185	0.91	4.6	150	x10
M	6.6	1	35	140	215	3	8.6	186	0.84	4.9	149	x11
M	6.4	1.5	41	129	220	2.75	8.5	175	0.79	4	136	x12
M	6.8	0.86	34	115	218	28	8.7	169	0.83	3.9	139	x13
M	7	0.9	37	150	222	2.6	9.3	182	0.81	4.2	146	x14
M	6.3	0.8	36	133	209	2.8	9.2	180	0.84	4	141	x15

sex	Uric acid	Serum creatinine	Blood Urea	Triglycerides	Total S. Cholestrol	Serum phosphorus	Serum calcium	Serum chloride	Serum Magnesium	Serum potassuim	Serum sodium	Patient
F	6.2	1.1	39	141	210	2.6	8.7	177	0.78	3.8	138	x16
F	5.9	1	36	120	195	3.6	8.1	167	0.81	3.9	141	x17
F	5.7	0.8	30	130	199	3.4	8.3	168	0.82	3.8	140	x18
F	6	0.9	33	152	211	3.2	8.6	182	0.91	4	144	x19
F	5.9	0.8	35	131	203	3.1	8.7	177	0.84	3.8	145	x20
F	6.2	0.7	32	110	198	3	8.8	173	0.85	4.4	142	x21
F	6	0.8	36	129	200	2.9	8.6	179	0.83	4.7	139	x22
F	5.6	0.9	37	145	219	3	8.8	184	0.82	4.2	141	x23
F	5.9	1	38	132	220	3.1	8.6	179	0.86	4.5	144	x24
F	5.5	0.8	34	100	190	3.5	8.3	168	0.81	4.3	147	x25
F	6.1	0.9	35	161	200	3	8.6	171	0.85	4.9	144	x26
F	6	0.9	39	133	206	2.9	8.7	169	0.82	4.8	148	x27
F	6.3	1.1	40	175	235	2.8	8.5	162	0.84	4.6	140	x28
F	5.2	1.5	44	130	209	2.8	8.7	167	0.88	4.2	144	x29
F	6.7	0.9	38	105	220	3.2	8.2	170	0.82	4.1	148	x30